DOCUMENT RESUME

ED 384 671 TM 023 992

| | |
|---|---|
| AUTHOR | Bridgeman, Brent; Rock, Donald A. |
| TITLE | Development and Evaluation of Computer-Administered Analytical Questions for the Graduate Record Examinations General Test. GRE Board Professional Report No. 88-06P. |
| INSTITUTION | Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program. |
| SPONS AGENCY | Graduate Record Examinations Board, Princeton, N.J. |
| REPORT NO | ETS-RR-92-49 |
| PUB DATE | Jan 93 |
| NOTE | 56p. |
| PUB TYPE | Reports - Research/Technical (143) -- Statistical Data (110) -- Tests/Evaluation Instruments (160) |
| EDRS PRICE | MF01/PC03 Plus Postage. |
| DESCRIPTORS | *Adaptive Testing; College Entrance Examinations; *College Students; *Computer Assisted Testing; Higher Education; Pattern Recognition; Tables (Data); *Test Construction; Test Items; *Thinking Skills |
| IDENTIFIERS | Free Response Test Items; *Graduate Record Examinations; *Open Ended Questions |

ABSTRACT

Three new computer-administered item types for the analytical scale of the Graduate Record Examination (GRE) General Test were developed and evaluated. One item type was a free-response version of the current analytical reasoning item type. The second item type was a somewhat constrained free-response version of the pattern identification (or number series) item type in which the student had to state the rule that generated the series. The third item type used the computer to administer yes/no analysis of explanation questions with a limited branching strategy. The computer tests were administered at four Educational Testing Service regional offices to a sample of students who had previously taken the GRE General Test. Scores from the regular GRE administration and the special computer administration were matched for a sample of 349 students. A number of test administration design issues were identified, including the need to provide adequate practice exercises, design of an interface comfortable for computer-literate students, and problems with item-level timing. The pattern identification items were too difficult (or the practice was inadequate), but the other items appeared to function well. There was no evidence that the open-ended analytical reasoning items were measuring anything beyond what is measured by the current multiple-choice version of these items. Eleven tables present study data. Four appendixes contain the various types of practice questions and the posttest questions. (Contains 23 references.) (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 384 671

GRE®

RESEARCH

# Development and Evaluation of Computer-Administered Analytical Questions for the Graduate Record Examinations General Test

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Brent Bridgeman
and
Donald A. Rock

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
E. COLEY
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 023 992

January 1993

GRE Board Professional Report No. 88-06P
ETS Research Report 92-49

BEST COPY AVAILABLE

Educational Testing Service, Princeton, New Jersey

# Development and Evaluation of Computer-Administered Analytical Questions for the Graduate Record Examinations General Test

Brent Bridgeman
and
Donald A. Rock

GRE Board Report No. 88-06P

January 1993

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Copyright © 1993 by Educational Testing Service. All rights reserved.

Acknowledgments

The authors wish to thank the test developers and researchers who contributed their ideas. Walter Emmerich, Mary Enright, and Carol Tucker shared their experiences with new multiple-choice items for the GRE analytical measure. Clark Chalifour made a number of useful observations on what might be gained, and lost, in computerizing the analytical reasoning items. Donald Powers suggested a strategy for computerizing the pattern identification items, and comments concerning other potential tasks were made by Timothy Habick, Spencer Swinton, William Ward, and Erich Woisetschlaeger.

Thanks to Caryn Ashare, Joyce Gant, Candus Hedberg, and the staff at the ETS regional offices in Austin, Evanston, Princeton, and Washington for their assistance with the data collection. The computer programs for the delivery and scoring of the test questions were ably written by Jeffrey Jenkins. Additional computerized scoring procedures were developed by Thomas Jirele. Study files were matched with GRE files by Patricia Lynn, and data analyses were efficiently performed by Inga Novatkowski and Minhwei Wang. Sabrina Waller typed the tables. Thanks to the GRE Research Committee for their interest in this project and their financial support.

Randy Bennett and Walter Emmerich provided a number of useful comments on an early draft of this paper.

Abstract

Three new computer-administered items types for the analytical scale of the Graduate Record Examination General Test were developed and evaluated. One item type was a free-response version of the current analytical reasoning item type. The second item type was a somewhat constrained free-response version of the pattern identification (or number series) item type in which the student had to state the rule that generated the series. The third item type used the computer to administer yes/no analysis of explanations questions with a limited branching strategy. The computer tests were administered at four ETS regional offices to a sample of students who had previously taken the GRE General Test. Scores from the regular GRE administration and the special computer administration were matched for a sample of 349 students. A number of test administration design issues were identified, including the need to provide adequate practice exercises, design of an interface comfortable for computer-literate students, and problems with item-level timing. The pattern identification items were too difficult (or the practice was inadequate), but the other items appeared to function well. There was no evidence that the open-ended analytical reasoning items were measuring anything beyond what is measured by the current multiple-choice version of these items.

## Introduction

The ability to reason critically and analytically is important both as an outcome of an undergraduate education and as a prerequisite for graduate training. In recognition of the importance of analytic thinking, the Graduate Record Examinations (GRE) General Test now includes an analytical measure. This score is derived from two item types: logical reasoning (LR) and analytical reasoning (AR). The LR type is actually a set of related types that are all based on one or occasionally two questions that follow a short passage in the form of an argument. The items assess such skills as recognizing assumptions, evaluating arguments and counterarguments, and analyzing evidence. The AR items present a brief scenario together with a set of rules relating different elements in the scenario; the four to seven questions on each scenario assess skills in combining rules to arrive at deductions on what must be true or could be true given the rules. Two additional item types that were developed for the analytical measure were dropped because they were susceptible to special test preparation and within-test practice effects (Kingston & Dorans, 1982; Swinton and Powers, 1983; Swinton, Wild, & Wallmark, 1983).

The construct validity of the current analytical measure is threatened because the two item types do not hold together as a separate construct that is distinct from the verbal (V) and quantitative (Q) measures; AR items correlate more highly with quantitative items than with LR items and LR items correlate more highly with verbal items than with AR items (Wilson, 1985). Factor analytic results using full information factor analysis (Schaeffer & Kingston, 1988) and confirmatory multidimensional item response theory (Kingston & McKinley, 1988) suggest that there is a weak analytic factor defined by the analytical reasoning items and not the logical reasoning items. Because the original analytical measure (with four item types) yielded a score that was more distinct from V and Q (Powers & Swinton, 1981), it is hoped that adding more item types to the current test will regain a distinctive analytic construct. Additional item types should also provide better coverage of the variety of reasoning skills that graduate faculty see as important for success in graduate school. In a study of graduate faculty in six fields of study, Powers and Enright (1987) identified a set of five dimensions that underlie faculty perceptions of reasoning ability: (a) the analysis and evaluation of arguments, (b) the drawing of inferences and the development of conclusions, (c) the definition and analysis of problems, (d) the ability to reason inductively, and (e) the generating of alternative explanations or hypotheses.

Emmerich, Enright, Rock, and Tucker (1991) developed and tested new analytical item types, including a revised version of analysis of explanations (one of the item types dropped from the original GRE analytical scale), numerical logical reasoning (based on the "ill-structured" multiple-choice questions developed by Ward, Carlson, and Woisetschlager [1983]), contrasting views (a variant of contrasting arguments [Carlton, 1987]), and pattern identification (number series with constraints). But these new items were necessarily constrained by the scannable-document multiple-choice (including multiple Yes/No) format.

The prospect of a computer-delivered GRE general test opens new possibilities for reasoning items. A wide range of different approaches are possible with computer delivery. For example, current or recent GRE-sponsored projects include using computers to deliver test questions in the following ways: (a) regular multiple-choice questions in a standard linear manner (phase one of the GRE General Test computerized delivery); (b) regular multiple-choice questions in an adaptive (branching) mode (the second phase of the GRE General Test project); (c) figural response questions in which the examinee rearranges material on the screen or uses a mouse to point to components of a figure displayed on the screen (Martinez, in preparation); (d) quantitative questions in which the examinee enters a numerical response (Bridgeman, 1991);(e) complex constructed response questions in which the computer evaluates solution strategies for quantitative questions (Sebrechts, Bennett, & Rock, 1991); and (f) questions that require the examinee to list (and the computer to score) hypotheses that could explain a result (Bennett & Kaplan, 1990). Most of these studies incorporate relatively well-defined scoring strategies that could be implemented in an operational test in the near future; the last two studies mentioned are exploring more experimental scoring algorithms that are not yet ready for routine operational use. The current project focused on reasoning questions that went beyond mere computerization of multiple-choice questions, ut that still had clear and defensible scoring rules that could be incorporated into an operational computer-delivered test in the short run.

Perhaps the most obvious advantage of a computerized reasoning test is that the examinee can be asked to generate answers rather than merely recognize them. There is evidence that free-response item types such as formulating hypotheses may tap different skills than their multiple-choice counterparts (Ward, Frederiksen, & Carlson, 1980). Furthermore, they may be better predictors of certain aspects of graduate education (Frederiksen & Ward, 1978).

The primary anticipated benefit of the new computer-delivered items was the possibility to more adequately assess the domain of reasoning skills with a concomitant potential for improved construct and criterion-related validity. But even in the absence of such benefits, the new items could be of value in substantially improving the face validity of the analytical score. In addition, free-response items might improve score accuracy by eliminating the effect of random guessing. Thus, the goals of the current project were to (1) identify new ways to assess analytical abilities, (2) identify advantages and obstacles created by computer delivery, and (3) determine whether the new questions help to define an analytical dimension for the GRE that is distinct from the verbal and quantitative dimensions.

## Method

### Item Development

Despite several productive meetings with GRE test development staff, consultations with inside and outside experts, and review of existing tests, the identification of items that met the constraints of the GRE program and were uniquely suited to computer administration proved to be a difficult task. As is the intent with the current items, the new items needed to be fair for men and women of all ethnic groups in all undergraduate majors; formal training in logic should not be required and should not provide a significant advantage if taken; and the new items should be reasonably independent of the existing verbal and quantitative reasoning dimensions. As one consultant noted, these constraints may define a null set.

Finding tasks that could not only be delivered on the computer but were also uniquely suited for this delivery mode was also difficult. One reason for this difficulty is that the multiple-choice format is actually quite adequate for many reasoning tasks where the number of plausible alternatives is quite limited. For example, a conclusion typically either follows or does not follow from a set of arguments, so a key list of 20 possible answers is not very useful. Three item types were developed for which computer administration did appear to be practical and to provide a potentially valuable new dimension. One of the new item types was derived from the current analytical reasoning item type, one was developed from the pattern identification item type proposed by Emmerich et al. (1991), and the third was derived from the analysis of explanations item type that was originally part of the GRE analytical score.

Analytical reasoning. For the analytical reasoning items, computer administration permits an item to ask the candidate to generate a solution that fits the rules rather than merely asking the candidate whether a provided solution conforms to the rules. Furthermore, computer administration allows assessment of cognitive flexibility by asking the candidate for more than one solution to a given problem. Three sets of four items each were developed from existing analytical reasoning scenarios, but no attempt was made to keep the new items parallel to the existing multiple-choice items. Indeed, the new items appear to be most useful to the extent that they cannot be made parallel to the old items.

In addition to the three scored problem sets, a fourth set of items was developed to provide practice in manipulating the computer interface. The text for the practice set and three scored sets of questions are in Appendix A, but note that the actual questions with their accompanying computer graphics look quite different from the text versions. One problem is a standard multiple-choice question (problem 4 in the first problem set). Some other problems (e.g., problems 2 and 3 in the second [airline] problem set) may at first glance appear to be standard multiple-choice questions but are not because of the multiple answer capability; there are 64 different possible answers to problem 3. The time limit for each problem was 3 minutes. A clock in the corner of the screen displayed the amount of time remaining for each

problem. To discourage impulsive responding, no bonus was awarded for answering quickly.

In one respect the computer version may be easier than the multiple-choice version. In the computer version, a pictorial representation of the problem is provided (e.g., the calendar for the first problem set or the seating chart for the second problem set) and the candidate can then manipulate symbols within that pictorial representation according to the rules stated. In the standard multiple-choice presentation only words are provided; any pictorial representation of those words must be provided by the candidate. If coding the information into a diagram (mentally or on paper) is considered an incidental task, the computer version may be more valid. But if the coding is considered part of the central construct, the paper-and- pencil version may be more valid.

Pattern identification. For these items, the candidate must generate the rule that relates the various numbers in the series, subject to constraints including the permissible operations (add, subtract, multiply, and divide) and the permissible numbers (1 to 4). For some problems, one rule is sufficient to generate all of the numbers in the series; for other problems, different rules may be needed for the even and odd members of the series; for still other problems, one rule may be needed for the numbers at the beginning of the series and a different rule for numbers later in the series. The multiple-choice version of this task (Emmerich et al., 1991) contained two additional possible series rules that were not included in the present task in an effort to simplify the instructions and the task itself.

For any given series, more than one rule may be correct (e.g., "multiply by 3 and then subtract 3" is equivalent to "subtract one and then multiply by three"). All potentially correct rules may not be anticipated by the item writer. Therefore, the computer scoring algorithm actually applies the rule generated by the candidate to the series rather than just checking against a prespecified answer key. A sample problem is in Appendix B. A maximum of 3 minutes was allowed for each item.

Analysis of explanations. A third item type that was evaluated is a computerized version of the analysis of explanations item type that was once a part of the analytical section of the general test. The original item type was dropped because it had very complicated directions that appeared to be especially susceptible to coaching. The computerized version leads the candidate through a series of yes or no decisions that essentially replicate the decision process with the old complicated directions. For each of the four problem sets, a fact situation is described in a paragraph of about 125 words followed by a one sentence result. Several statements follow the result. For each statement, the candidate must first indicate whether the statement is inconsistent with anything in the fact situation or result. If it is not inconsistent, the candidate must then decide if the statement is deducible from the fact situation and/or the result. If it is neither inconsistent nor deducible, the candidate must then decide if the statement is relevant to a possible explanation of the result. If a candidate made an error at the first level, the later levels were not administered and were automatically scored as incorrect. For example, if a statement were relevant

to a possible explanation but the candidate indicated that the statement was inconsistent with the fact situation, neither the "deducible" nor the "relevant" questions would be asked, and the candidate would automatically get a 0 score for those two questions. For 5 of the 32 questions, the statement was inconsistent with the fact situation and/or result, and for 7 questions the statement was deducible. For the remaining 20 questions, the candidate had to make a judgment on relevance. Note that three separate yes or no decisions were required for each of these 20 questions. Thus, a total of 79 yes or no decisions were evaluated with the 32 statements[1]. (See Appendix C.) A free-response section at the end allowed the candidate to enter a plausible explanation that the item developers did not consider. These free responses will initially have to be evaluated by human judges but may eventually be computer scorable. Three minutes were allowed to read the situation and the result and to answer the first question in each problem set. Subsequent items within each set had a 1-minute time limit.

Subjects and Field Trial Procedures

The new reasoning items, together with 14 new open-ended quantitative items that were developed in a separate parallel project (Bridgeman, 1991)[2]; were evaluated in a field trial at the end of February, 1990. Students who had taken the October 1989 GRE General Test, who had completed the biographical information questionnaire (BIQ), and who lived near one of the four Educational Testing Service (ETS) offices where the computer test was to be administered (Austin TX, Evanston IL, Princeton NJ, and Washington DC) were sent letters inviting them to participate in a study "designed to evaluate some new computer-delivered test items that have been developed for the GRE General Test." They were told that they would be paid $40 for the 2-hour testing session that was to take place at the local ETS regional office. Invitation letters were sent to 3,277 candidates. They were told that a limited number of testing appointments were available and would be filled on a first-called, first-scheduled basis. The available slots in the four centers filled within a few weeks of the mailing. A total sample of 364 candidates was eventually tested. Although this sample was geographically diverse and represented a range of skills and background characteristics, it should not be considered as a random sample because of its volunteer nature. Thus, for example, candidates who were particularly apprehensive about taking a computer-administered test may be underrepresented.

Sample description

Of the 364 candidates who took the computer test, 15 could not be matched to the data base of BIQ and regular GRE scores because they failed to

---

[1]Although this description may appear to be quite complicated, the task faced by the examinee was very simple. At any given time, the examinee had to make a single yes or no answer to the question presented on the screen.

[2]These 14 items were derived from regular GRE quantitative items. The answer choices were removed and the examinees were asked to enter numerical answers with the computer keyboard.

enter valid identification numbers. Regular GRE scores for the remaining 349 candidates are summarized at the bottom of Table 1. For comparison purposes, the scores of the total population of GRE test takers are presented at the top of the table; these scores from the 1987 - 1988 test year are the most recent now available (Wah & Robinson, 1990). Although the experimental sample was not limited to citizens of the United States, all of the test centers for the

Table 1

Comparison of GRE Scores for GRE Population and Experimental Sample

| | | GRE Scores | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Verbal | | | Quantitative | | | Analytical | | |
| Group | n | M | SD | d | M | SD | d | M | SD | d |
| Population[a] | | | | | | | | | | |
| Total | 278,878 | 486 | 122 | | 553 | 139 | | 529 | 128 | |
| Men | 134,469 | 484 | 128 | | 599 | 135 | | 535 | 132 | |
| Women | 144,369 | 487 | 117 | | 510 | 129 | | 524 | 124 | |
| Cender Difference | | | | -.02 | | | .67 | | | .09 |
| U.S. Citizens[a] | | | | | | | | | | |
| Total | 221,638 | 508 | 114 | | 537 | 135 | | 542 | 125 | |
| Men | 95,142 | 519 | 116 | | 583 | 134 | | 557 | 127 | |
| Women | 126,496 | 499 | 113 | | 502 | 125 | | 531 | 123 | |
| Gender Difference | | | | .17 | | | .63 | | | .21 |
| Experimental | | | | | | | | | | |
| Total | 349 | 544 | 120 | | 578 | 129 | | 584 | 125 | |
| Men | 131 | 557 | 126 | | 611 | 134 | | 591 | 130 | |
| Women | 218 | 536 | 115 | | 558 | 122 | | 580 | 122 | |
| Gender Difference | | | | .17 | | | .42 | | | .09 |

[a]Population and U.S. citizens are from 1987 - 1988 test year (Wah & Robinson, 1990).

computer administration were in the United States, and therefore U.S. citizens are overrepresented compared to the GRE population as a whole. About 21% of the GRE population is composed of non citizens, but only 7% of the experimental group were non citizens. Table 1 also presents the means and standard deviations for U.S. citizens. Mean scores of the self-selected experimental group were about one-third of a standard deviation higher than the mean scores for the U.S. citizens. The experimental group was quite heterogeneous, with standard deviations approximately equivalent to the standard deviations in the entire GRE population. Gender differences are indicated in the table by d; which is the mean difference divided by the pooled within-gender standard deviation; positive values indicate higher mean scores for men. Consistent with the differences noted for U.S. citizens, men in the experimental group received higher scores than women on all three score scales. However, for both the quantitative and analytical scales the gender difference was slightly smaller than would be expected based on the U.S. citizens sample.

Although the degree of computer literacy in the overall GRE population was unknown, the experimental sample was generally familiar with computers. On the questionnaire that was administered at the end of the testing session (see Appendix D), 92% of the sample reported using the computer for word processing at least once in the last two years; 79% reported using it at least five times. For uses other than word processing, 83% of the sample reported using the computer at least once in the past two years; 55% used it five or more times.

## Analyses

Analyses included basic descriptive statistics on the experimental measures, including means, standard deviations, and reliabilities. Responses to the questionnaires that the examinees completed at the end of the testing session were analyzed to determine perceived task difficulty and to identify problems with test directions or timing. Relationships among the current and experimental item types were explored with correlational analyses, and with correlations corrected for unreliability.

The relationships among the multiple-choice and computer items were further explored with exploratory and confirmatory factor analyses. In order to better approximate the linear factor model assumption of multivariate normality, item parcels rather than individual items were analyzed. Each parcel was constructed from a minimum of four items.

Parcel definition. Each of the four item types on the regular verbal scale (sentence completion, analogies, reading comprehension, and antonyms) was divided into two parcels by an odd-even split, yielding eight markers for the verbal factor. Similarly, each of the three item types on the quantitative scale (quantitative comparisons, discrete quantitative, and data interpretation) yielded six markers for the quantitative factor. An additional two parcels were created from an odd-even split of the computer quantitative items (Bridgeman, 1991). The two item types on the analytical scale (analytical reasoning and logical reasoning) were divided into three parcels each in order to generate a sufficient number of markers for an

analytical factor. The three analytical reasoning parcels were created in a manner such that all questions based on a single problem statement were assigned to the same parcel. A similar constraint was imposed for the creation of three parcels from the computerized analytical reasoning items and of two parcels from the analysis of explanations items. Because of the overlap across scales on analysis of explanations, parcels were created for only the 20 items on the "relevant" scale. Pattern identification items were divided into two parcels by an odd-even split. Parcels within an item type were inspected to ensure that the parcels were roughly equivalent in terms of mean difficulty; no adjustments were necessary. The above procedures resulted in the creation of 29 parcels. Defining parcels within item types allows the emergence of the maximum number of factors. More parsimonious models can then be created by inspection of the factor intercorrelations.

Because each examinee in the experimental sample had taken one of the four different versions of the GRE General Test that was administered in October 1989, parcel definition and intercorrelations across parcels were conducted separately for each form. The four variance-covariance matrices were then averaged to provide the single matrix that was factor analyzed.

It is important to remember that the computer test parcels differed from the multiple-choice parcels not only in mode of delivery but also in time and in the motivational level of the students taking them. The computer tests were administered four months after the multiple-choice tests, and students knew that scores on the computer tests would not be reported to graduate schools (or anyone else except the researchers).

Exploratory factor analysis. The correlation matrix of item parcels was first analyzed with a principal components model (with 1s on the diagonal) to determine the number of factors to extract. Then the matrix was factor analyzed with Minres (communalities [squared multiple correlations] on the diagonal) and rotated with Promax (which allows correlated factors).

Confirmatory factor analyses. The variance covariance matrix of parcels was evaluated with the EQS structural equation program (Bentler, 1985) using maximum likelihood factor estimation procedures (Jöreskog & Sörbum, 1985). Alternative models were estimated based on different assumptions with respect to both the number of factors and the general pattern of trivial and salient loadings. Strictly speaking, only the eight-factor model represented a truly "confirmatory" factor analysis in that this model was the only one fully specified before the data were examined. Nevertheless, the techniques of confirmatory factor analysis are very helpful for describing the characteristics of more parsimonious models. These competing models were then tested for goodness of fit to the data. Because a universally accepted measure of fit does not exist for confirmatory factor analysis (Marsh & Hocevar, 1985), several different measures that are sensitive to different departures in fit were employed.

The goodness-of-fit indicators used included the Tucker-Lewis (T-L) index (Tucker & Lewis, 1973), the chi-square/degrees of freedom ($X^2/df$) ratio, and the mean off-diagonal standardized residual. The T-L index represents the ratio of the amount of variance associated with the model to the total

variance, and may be interpreted as indicating how well a factor model with a given number of common factors represents the covariances among the parcels for a population of examinees. It may be interpreted as a reliability coefficient, with low values indicating that the relations among the parcels are more complex than can be represented with that number of common factors. The $X^2/df$ ratio is based on the overall goodness-of-fit test for each model. Because the ratio depends on sample size, it must be interpreted cautiously. (This is not of great concern for the current study because all models were run on the same sample.) Ratios up to about 5.0 indicate a reasonable fit (Marsh & Hocevar, 1985).

## Results and Discussion

In this section, the characteristics of the tasks as new measures of analytical abilities are discussed first. Next, issues related to the use of the computer as a testing device are discussed. Finally, the contributions of the new tasks to the creation of a distinct analytical factor are presented.

### Characteristics of Experimental Measures

Means and standard deviations for the two analytical item types on the regular GRE general test (analytical reasoning and logical reasoning) are presented in Table 2 along with the means and SDs for the computer tests and the gender differences in d units. For both of the standard multiple-choice scales the mean scores were above the midpoint of the raw score scale.

Analytical reasoning. The mean for the computer-delivered analytical reasoning scale represented getting about half of the items correct. The questionnaire responses on perceived difficulty were consistent with this indication of actual difficulty. About 65% of the sample indicated that the difficulty was "about right," with 22% indicating that it was too easy and 13% indicating that it was too difficult. Similarly, time limits were perceived as appropriate (65% about right, 27% too short, and 9% too long). The directions for this section were seen as easy by 86% of the sample and difficult by 14%. Although the women scored slightly higher on average than the men, the difference was neither practically nor statistically significant.

Pattern identification. The mean score for the pattern identification items was only 2.5 out of a possible 8. Figure 1 presents the percentage of each gender group in each score category. In both gender groups the mode (i.e., most frequent score) was 0. For women, the next most frequent score was 1; for men, roughly equal numbers were in each score category from 1 through 6. The gender difference in mean scores was statistically significant, $t(347) = 2.83$, $p < .01$.

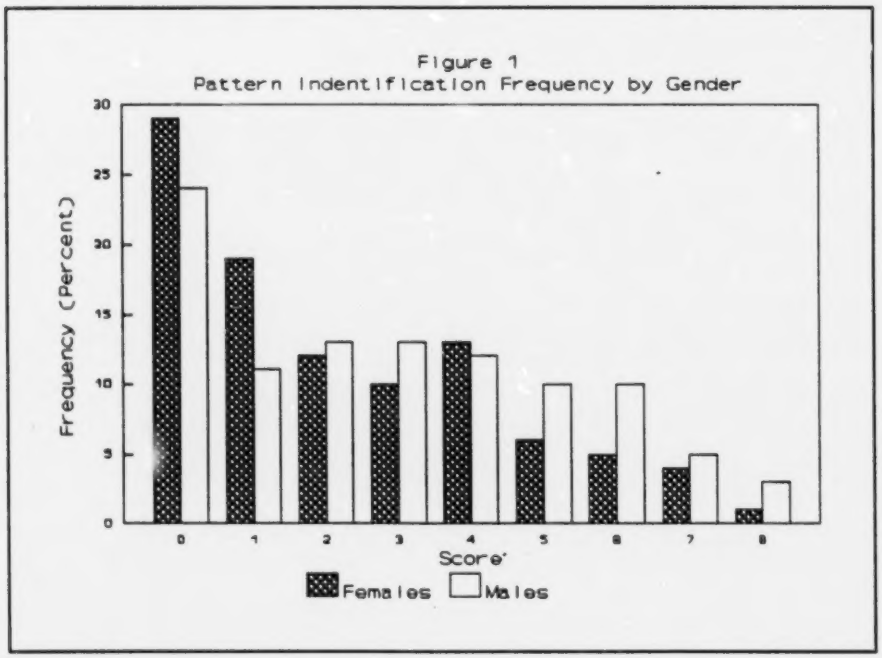
Slightly over half of the sample (52%) reported that the pattern identification test was too difficult; 44% indicated that it was about right, and 5% thought it was too easy. The time limit of three minutes per problem was seen as too short by 53% of the sample; 43% indicated that the time limit was about right, and only 4% thought it was too long. The directions were perceived as hard to understand by 49% of the sample.

Table 2

Means and SDs for Standard and Computer-Administered Analytical Reasoning Scores

| | | Standard Multiple-Choice | | | | | | | | Computer Delivered | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Analytical Reasoning | | | | Logical Reasoning | | | | Analytical Reasoning | | | | Pattern Identification | | | | Analysis of Explanations | | | |
| Gender | n | I[a] | M | SD | d | I | M | SD | d | I | M | SD | d | I | M | SD | d | I | M | SD | d |
| Total | 349 | 38 | 24.5 | 7.2 | | 12 | 7.8 | 2.5 | | 12 | 6.3 | 2.7 | | 8 | 2.5 | 2.2 | | 79 | 48.4 | 17.3 | |
| Men | 131 | | 24.7 | 7.4 | | | 8.1 | 2.6 | | | 6.1 | 2.8 | | | 2.9 | 2.3 | | | 49.4 | 17.8 | |
| Women | 218 | | 24.4 | 7.0 | | | 7.7 | 2.4 | | | 6.4 | 2.7 | | | 2.2 | 2.2 | | | 47.9 | 17.2 | |
| Gender Difference | | | | | .06 | | | | .16 | | | | -.11 | | | | .30* | | | | .09 |

[a]"I" is the number of items on the scale
*$p < .01$




Analysis of explanations. The mean score of 48.4 out of a possible 79 (or 61%) on the analysis of explanations test may at first appear to be quite low on a test that consists of only yes and no questions. However, because of the branching administration design a random guesser would get considerably fewer than half of the 79 items correct. For example, if the explanation provided is relevant, the examinee must first indicate that it is not inconsistent and then indicate that it is not deducible from the information given; only then is the examinee asked to make a judgment on relevance. An incorrect guess on the first part of the item means that the examinee will not be given a chance to guess on the next two parts of the item and will receive an automatic 0 on those parts.

Questionnaire responses suggested that the difficulty level of this test was appropriate (61% about right, 19% too easy, and 20% too difficult). Time limits were perceived as too short by 35% of the sample, about right by 60% and too long by 5%. The directions were rated easy to understand by 68% of the sample.

Means and standard deviations for the three analysis of explanations subscores are presented in Table 3. As indicated above, each of the 32 questions on the "inconsistent" subscale reflected a single yes or no question; each of the 27 questions on the "deducible" subscale required correct answers to two yes or no questions; and each of the 20 questions on the "relevant" subscale required the correct answer to three yes or no questions. Thus, on average, a group of random guessers would get 16 on the inconsistent subscale, 6.75 on the deducible subscale, and only 2.5 on the relevant subscale. Gender differences were small ($ds<.2$) and not statistically significant ($ps>.05$).

Table 3

Means and Standard Deviations for Analysis of Explanations Subscores

| Gender | Analysis of Explanation Subscores | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Inconsistent | | | | Deducible | | | | Relevant | | | |
| | I | M | SD | d | I | M | SD | d | I | M | SD | d |
| Total | 32 | 23.4 | 5.8 | | 27 | 15.8 | 6.7 | | 20 | 9.3 | 5.6 | |
| Men | | 23.3 | 5.8 | | | 16.2 | 6.9 | | | 9.9 | 5.8 | |
| Women | | 23.4 | 5.8 | | | 15.5 | 6.6 | | | 8.9 | 5.5 | |
| Gender Difference | | | | -.02 | | | | .11 | | | | .17 |

Reliability

Coefficient alpha reliability estimates for the multiple-choice and experimental analytical scores are presented in Table 4. Because experimental subjects took one of the four regular GRE forms that were used for the October 1989 test administration, reliabilities for the multiple-choice items were computed within form and then averaged across forms. To avoid inflated reliability estimates that theoretically can result when related items are considered together (e.g., when several items relate to a single problem statement), analytical reasoning and analysis of explanations items were grouped into homogeneous parcels before alpha was computed. For example, the computer analytical reasoning test consisted of three problem statements with four questions related to each statement. Each problem statement with its four related questions formed a parcel score with its own mean and variance; alpha was computed from the relationships among these three parcel scores.

Comparisons among reliability coefficients that are based on tests of different lengths must be made cautiously. Between the two multiple-choice scores on the analytical scale, analytical reasoning appears to be considerably more reliable than logical reasoning. But the analytical scale contains 38 analytical reasoning items and only 12 logical reasoning items. The next to the last column of Table 4 uses a Spearman-Brown adjustment to estimate the reliability of a 30-item test. For a constant-length test, the previous picture reverses, and the logical reasoning items now appear to be more reliable than the analytical reasoning items. With the adjusted reliability coefficients, the experimental scales all compare favorably with the existing multiple-choice scores. However, the constant-length criterion may overstate the reliability of the open-ended tests relative to the multiple-choice tests because the open-ended tests typically require more time per item. (Note the "estimated seconds per item" column in Table 4.) A 30-item multiple-choice analytical reasoning test would require about 36 minutes

Table 4

Observed and Estimated Alpha Reliabilities for Standard and Computer Tests

| Test | Number of Items | Seconds per Item | Observed Alpha | Alpha for 30-items | Alpha for 30-minutes |
|---|---|---|---|---|---|
| **Multiple-Choice** | | | | | |
| Analytical Reasoning | 38 | 72 | .80 | .76 | .72 |
| Logical Reasoning | 12 | 72 | .69 | .85 | .82 |
| **Computer Delivered** | | | | | |
| Analytical Reasoning (Open-ended) | 12 | 180 | .76 | .89 | .72 |
| Pattern Identification (Open-ended) | 8 | 180 | .84 | .95 | .87 |
| Analysis of Explanations--Relevant scale (yes/no) | 20 | 66 | .89 | .92 | .92 |

(with current timing standards), which is the same as the time needed for the 12-item computer test (not counting the time needed to read the directions and try the practice exercises). Alpha reliabilities for a 30-minute test, as estimated by a Spearman-Brown adjustment, are presented in the last column of Table 4. Although there is no assurance that the timing of the computer test was optimal, recall that 65% of the students in the sample thought the timing was about right and 27% thought it was too short. The elimination of random guessing undoubtedly has a positive impact on reliability, but it may be offset by the greater time required for the free-response questions. The computer-delivered yes/no questions in the analysis of explanations test were both highly reliable and relatively efficient on a time-per-question basis.

## Computer Delivery Design Considerations

Interviews with students conducted immediately after they completed the computer tests suggested that a few modifications in the system might make it considerably easier to use. In an attempt to make the system easy to use by students with little or no computer experience, we unintentionally made it difficult for computer-literate students. We reasoned that even students with no computer experience could easily find the space bar, so the space bar was used for screen navigation (e.g., use the space bar to move the highlight box to next part of the diagram). The problem was that this was a very unnatural way to move around on the screen for students who were even minimally computer literate, which turned out to be almost everyone in the sample. The students indicated that they would have preferred a mouse or at least the use of the arrow keys for this type of screen navigation.

Timing individual items also caused a problem. Some kind of time constraint was needed to eliminate endless trial-and-error strategies that could eventually lead to the right answer even among students who had a poor conceptual understanding of the problem. Because we were concerned that students would not know how to pace themselves on the totally unfamiliar computer-administered item types, time limits were enforced on individual items. But imposing a time constraint on individual items meant that a student who was just a little slower than average might get no items correct. The frustration at running out of time to complete the last few items in a section does not compare to the frustration at running out of time on every single item. In the former case, the total score might be a few points lower for the slow but accurate student; in the latter case the same slow student could end up with a score of 0.

Although on-line practice exercises were generally seen as a good feature, some students complained that the practice was inadequate for question types that were quite unlike anything they had seen before. In the current study, the examinees received no practice materials ahead of time. Expanding the practice exercises at the beginning of the testing session might help to alleviate this problem, but the real solution is probably to provide explanations and computerized practice materials several weeks before the test administration. Advanced practice on these materials would also give students an opportunity to learn how to pace themselves so that section timing could replace individual item timing.

## Distinctiveness of New Analytical Scores

Correlations among scores. Correlations among the various scores are presented above the diagonal in Table 5 and correlations corrected for unreliability below the diagonal. Even after correction for unreliability, the highest correlations for the pattern identification test were in the .60s. Thus, the current result is consistent with the findings for the multiple-choice version of this question type, suggesting that it contains substantial reliable variance that is not shared with other reasoning tests (Emmerich et al., 1991). Corrected correlations of pattern identification with multiple-choice quantitative and analytical reasoning scores from the current sample were almost identical to the correlations in the Emmerich et al. (1991)

Table 5

Corrected and Uncorrected Correlations Among Scores

| Score | V-MC | Q | | A-MC | LR-MC | AR | | PI-C | AE-YN | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MC | O | | | MC | O | | tot | I | D | R |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1. Verbal-MC | | .51 | .44 | .65 | .70 | .54 | .44 | .35 | .62 | .59 | .59 | .62 |
| 2. Quantitative-MC | .51 | | .77 | .73 | .52 | .72 | .55 | .57 | .49 | .45 | .47 | .49 |
| 3. Quantitative-O | .56 | .91 | | .65 | .46 | .63 | .55 | .57 | .51 | .48 | .49 | .51 |
| 4. Analytical-MC | .74 | .84 | .78 | | .72[a] | .97[a] | .71 | .54 | .64 | .61 | .62 | .61 |
| 5. Logical reasoning-MC | .88 | .65 | .62 | .94[a] | | .53 | .46 | .32 | .56 | .53 | .54 | .53 |
| 6. Analytical reasoning-MC | .63 | .84 | .79 | 1.18[a] | .71 | | .70 | .55 | .59 | .56 | .56 | .56 |
| 7. Analytical reasoning-O | .56 | .69 | .73 | .92 | .66 | .94 | | .47 | .59 | .56 | .57 | .56 |
| 8. Pattern indentification-O | .40 | .65 | .69 | .64 | .42 | .68 | .61 | | .46 | .44 | .44 | .43 |
| Analysis of Explanations-YN | | | | | | | | | | | | |
| 9. Total | .67 | .53 | .59 | .71 | .69 | .67 | .72 | .51 | | .94[a] | .98[a] | .95[a] |
| 10. Inconsistent | .64 | .49 | .55 | .69 | .66 | .65 | .69 | .50 | 1.00[a] | | .88[a] | .81[b] |
| 11. Deductible | .64 | .52 | .57 | .70 | .68 | .66 | .71 | .50 | 1.05[a] | .96 | | |
| 12. Relevant | .70 | .56 | .61 | .71 | .68 | .67 | .72 | .50 | 1.04[a] | .90[b] | 1.05[b] | .92[b] |

Note: Correlations corrected for attenuation below diagional, uncorrected above
MC-Traditional multiple-choice scores from regular GRE administration
O-Open-ended scores from computer test
YN-Yes/No scores from computer test
[a]part-whole correlation
[b]some dependence between scores

sample (.65 and .68, respectively, in the current sample and .64 and .64 in the previous sample). However, the multiple-choice version of pattern identification was more highly correlated with GRE verbal and logical reasoning (.50 and .63, respectively) than was the computer version of pattern identification (.40 and .42, respectively). This may reflect the greater verbal load in the multiple-choice version related to the need to comprehend two and a half pages of written directions before beginning the problems. Although the directions for the computer version were also complex, the learning task was divided into small pieces with hands-on practice provided for each step.

With the unreliability correction, the multiple-choice and computer versions of analytical reasoning correlated .94, suggesting that to a considerable extent they are both measures of the same underlying construct. Corrected correlations with GRE verbal, logical reasoning, and analysis of explanations were approximately the same for the multiple-choice and computer versions of the analytical reasoning score. However, the correlations with GRE quantitative were quite discrepant (.84 for the multiple-choice test and .69 for the computer version). This may be related to the problem of pictorial representation discussed above. The ability to turn words into pictures may be important for success on both quantitative and multiple-choice analytical reasoning questions, but this skill is not needed for the computer analytical reasoning items. It might be possible to make the existing multiple-choice analytical reasoning more independent of quantitative ability by providing a diagram in the test booklet.

In terms of correlations with other scores, the analysis of explanations subscores appear to be interchangeable. Analysis of explanations has about equal correlations (in the .67 to .72 range) with GRE verbal, analytical reasoning (computer or multiple-choice), and logical reasoning. This is in contrast to the logical reasoning items that have a substantially higher correlation with GRE verbal (.88) than with analytical reasoning (.71 and .66 for the multiple-choice and computer scores, respectively). The analysis of explanations items are also relatively independent of GRE quantitative, with a corrected correlation of .56 (or .61 with the computer-delivered quantitative items).

Exploratory factor analyses. In the principal components analysis, three components had eigenvalues greater than one, but the next two components were very close (.98 and .96). This was followed by a major drop to .83 with a relatively even drop beyond this point. Thus, the screen test suggested a five-factor model.

Results of the factor analysis (with communalities on the diagonal) and Promax rotation are presented in Table 6; only loadings of at least .30 are included. The solution was extremely clean in that no parcel loaded on more than one factor and the two or three separate parcels for each item type always loaded on the same factor. In addition, the factors were easily interpretable. Factor 1 contained all of the item types from the verbal scale plus the logical reasoning items. Factor 2 contained all of the item types from the quantitative scale plus the computer-administered quantitative items. Factor 3 contained only the data interpretation parcels, and Factor 4

Table 6

Factor Loadings for the Exploratory Factor Analysis

| Parcels | | Factors 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Sentence Completion | A | .66 | | | | |
| | B | .66 | | | | |
| Analogies | A | .72 | | | | |
| | B | .69 | | | | |
| Reading Comprehenison | A | .63 | | | | |
| | B | .62 | | | | |
| Antonyms | A | .80 | | | | |
| | B | .77 | | | | |
| Logical Reasoning | A | .50 | | | | |
| | B | .43 | | | | |
| | C | .53 | | | | |
| Quantitative Comparing | A | | .81 | | | |
| | B | | .81 | | | |
| Discrete Quantitative | A | | .78 | | | |
| | B | | .84 | | | |
| Data Interpretation | A | | .57 | | | |
| | B | | .62 | | | |
| Computer Quantitative | A | | .64 | | | |
| | B | | .62 | | | |
| Pattern Indentification | A | | | .91 | | |
| | B | | | .73 | | |
| Analysis of Explanation | A | | | | .73 | |
| | B | | | | .80 | |
| Analytical Reasoning | A | | | | | .44 |
| | B | | | | | .44 |
| | C | | | | | .32 |
| Computer-administered Analytical Reasoning | A | | | | | .48 |
| | B | | | | | .42 |
| | C | | | | | .48 |

Note: Loadings less than .30 omitted

contained only the analysis of explanations parcels. Factor 5 included analytical reasoning parcels from both the multiple-choice and open-ended versions of the task. Note that in the only two cases where multiple-choice and open-ended versions of the same task were available (i.e., quantitative and analytical reasoning), both formats loaded on the same factor and no distinct method factor emerged. Correlations among the factors are presented in Table 7.

Table 7

Correlations Among Factors for the Exploratory Model

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. Verbal + Logical Reasoning | -- | | | | |
| 2. Quantitative | .57 | -- | | | |
| 3. Pattern Identification | .38 | .59 | -- | | |
| 4. Analysis of Explanations | .55 | .45 | .38 | -- | |
| 5. Analytical Reasoning | .41 | .47 | .42 | .40 | -- |

Confirmatory factor analyses, eight-factor model. The first model tested was intended to represent a reasonable maximum for the number of factors. The parcels for the regular multiple-choice items were divided into four groups representing a verbal factor, a quantitative factor, and two factors from the analytical portion of the test. Previous studies (e.g., Rock, Bennett, & Jirele, 1988) suggested that a better fit is obtained when analytical reasoning items are separated from logical reasoning items than when they are grouped together as a single factor. Each of the computer tests (including the computer quantitative test) was included as a separate factor. Tests were constrained to load on only one factor each.

Goodness-of-fit statistics for the eight-factor model and the other models are presented in Table 8. As should be expected with a large number of factors, the eight-factor model fit very well. Of more interest are the correlations among the factors that suggest how factors could be most reasonably combined.

The correlations among factors are presented in Table 9 along with the correlation of each factor with undergraduate grade point average. Standard errors for the correlations among factors ranged from .03 to .06; all correlations for this and the other models were significantly different from 0 and from 1 ($p < .05$).

Table 8

Goodness of Fit Indicators for Alternative Models

| Model | $X^2/df$ | Tucker-Lewis | Mean Off-diagonal Standardized Residual |
|---|---|---|---|
| Eight-factor | 1.5 | .97 | .029 |
| Five-factor | 1.7 | .95 | .032 |
| Three-factor (LR with V) | 2.4 | .91 | .037 |
| Three-factor (LR with AR and AE) | 2.7 | .90 | .044 |
| Two-factor | 3.1 | .87 | .048 |

The correlations among factors generally confirm the corrected correlations in Table 5 except that the correlation between the verbal factor and multiple-choice analytical reasoning factor (r=.80) was higher than expected. It was also high in comparison to those found in other studies (.68 in the four-factor solution in Rock et al., 1988, and .58 in the corrected test correlations of Emmerich et al., 1991). Of particular importance for the development of a model with fewer factors was the correlation of .90 between the verbal and logical reasoning factors. This was consistent with both Rock et al. (1988) and Emmerich et al. (1991), which found correlations of .86 and .88, respectively. Thus, the logical reasoning items could probably be placed on the verbal factor with only minimal loss of fit. It also appeared that the computer and multiple-choice quantitative items could be placed on the same factor (r=.90). Similarly, the multiple-choice and computer analytical reasoning items were largely redundant (r=.93).

Five-factor model. The five-factor model was created from the eight-factor model by placing the logical reasoning parcels on the verbal scale, combining the multiple-choice and computer quantitative parcels, and combining the multiple-choice and computer analytical reasoning parcels. This five-factor model was also consistent with the results of the exploratory factor analysis.

As indicated in Table 8, the five-factor model fit the data nearly as well as the eight-factor model. The correlations among the factors presented in Table 10 suggest that further combining might result in a substantially poorer fit; the highest correlation among factors was .80. Analysis of explanations was about equally correlated with the verbal (including logical reasoning) factor and the analytical reasoning factor, but it was clearly

Table 9

Correlations Among 8 Factors and Factors with UGPA

| Score | V-MC | Q | | LR-MC | AR | | PI-C | AE-YN |
|---|---|---|---|---|---|---|---|---|
| | | MC | O | | MC | O | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1. Verbal-MC | -- | | | | | | | |
| 2. Quantitative-MC | .66 | -- | | | | | | |
| 3. Quantitative-O | .59 | .90 | | | | | | |
| 4. Logical Reasoning-MC | .90 | .63 | .62 | | | | | |
| 5. Analytical Reasoning-MC | .80 | .83 | .78 | .70 | | | | |
| 6. Analytical Reasoning-O | .66 | .67 | .71 | .65 | .93 | | | |
| 7. Pattern Indentification-O | .47 | .65 | .69 | .42 | .68 | .60 | | |
| 8. Analysis of Explanations-YN | .70 | .54 | .60 | .68 | .67 | .71 | .50 | |
| UGPA[a] | .27 | .22 | .23 | .23 | .32 | .26 | .14 | .25 |

Note: MC-multiple-choice, O-Open-ended, YN-Yes/No
[a]Undergraduate grade point average

Table 10

Correlations Among Factors for the Five Factor Model

| | V+LR 1 | Q 2 | AR 3 | PI 4 | AE 5 |
|---|---|---|---|---|---|
| 1. Verbal + Logical Reasoning | -- | | | | |
| 2. Quantitative | .66 | -- | | | |
| 3. Analytical Reasoning | .77 | .80 | | | |
| 4. Pattern Identification | .47 | .67 | .66 | -- | |
| 5. Analysis of Explanations | .71 | .56 | .70 | .50 | -- |

reliably measuring something that was unique. Pattern identification was also distinct (highest correlation was .67 with the quantitative factor).

Although the factor intercorrelations suggested that the three proposed analytical measures (analytical reasoning, pattern identification, and analysis of explanations) each represented a distinct dimension, a reduced model that more nearly matched the existing three General Test scores was also of interest. Even in a reduced model, pattern identification would require a separate factor because of its low correlation with any other variable. Given the other problems with this test already noted (e.g., modal score of 0 and low correlation with undergraduate grade point average [Table 9]), it was dropped from consideration for the three-factor models.

Three-factor models. Two three-factor models were evaluated. In one model, the logical reasoning parcels were combined with the verbal parcels to form one factor; in the other model, the logical reasoning parcels were included with the other analytical tests (analytical reasoning and analysis of explanations) because the logical reasoning items currently are included in the analytical score. In both models, the quantitative parcels (both computer and multiple-choice) were placed on a single quantitative factor. As indicated in Table 8, the fit statistics for the model with the logical reasoning items on the verbal factor were not substantially higher than the statistics for the alternative three-factor model.

The correlations among factors in the three-factor models are presented in Tables 11 and 12. The second model might be preferred because the factors appear to be more discriminable in that model. However, until these results can be replicated, little weight should be given to the small differences between these models. It would appear that either model describes the

relationships among the tests reasonably well, although the fit is noticeably worse than for the five-factor model.

Table 11

Correlations Among Factors for the Three Factor Model (LR with V)

| | V+LR<br>1 | Q<br>2 | AR+AE<br>3 |
|---|---|---|---|
| 1. V+LR | -- | | |
| 2. Quantitative | .65 | -- | |
| 3. AR+AE | .86 | .80 | -- |

Table 12

Correlations Among Factors for the Three Factor Model (LR with AR and AE)

| | V<br>1 | Q<br>2 | AR+LR+AE<br>3 |
|---|---|---|---|
| 1. Verbal | -- | | |
| 2. Quantitative | .66 | -- | |
| 3. AR+LR+AE | .81 | .79 | -- |

Two-factor model. As indicated in Table 8, the two-factor model (with one factor containing quantitative and analytical reasoning parcels and the other factor containing verbal, logical reasoning, and analysis of explanations parcels) did not fit as well as the other models, but the drop from the three-factor models was not precipitous. The correlation between the two factors was .74.

## Conclusions

The research presented here and in the other projects currently underway represent only the first steps in learning to take full advantage of the computer as an enabling technology for GRE testing. Nevertheless, with the exception of the pattern identification items, it appears that the computer tests developed in this project could be incorporated in a computer-delivered GRE General Test with relatively little difficulty. The delivery and scoring

mechanisms appeared to work as they were designed to function, and the items had reasonable psychometric characteristics.

Few conclusions concerning the possible utility of pattern identification items can be drawn from the current study. The mean score was only 2.5 out of a possible 8, with a modal response of 0. Despite a lengthy practice exercise, nearly half of the students thought the directions were hard to understand. Although the directions and practice exercises could undoubtedly be improved, a large part of the problem appears to be that the instructions are inherently complex. In the 3-option multiple-choice version of this task, Emmerich et al. (1991) noted a small but statistically significant practice effect on the test items even though the examinees completed a six-minute practice session before beginning the test. The potential value of the task in either the multiple-choice or computer format must be weighed against the lengthy instruction and practice session that appears to be necessary.

For the analysis of explanations test, the computer was not used to administer open-ended items. Rather, the computer simply provided a means to administer a branching series of yes and no questions. A task that had very complicated directions in a multiple-choice format thus became much easier for the student to follow. Psychometrically, the task appeared to perform very well; reliability was high, gender differences were small, and correlations with other scores suggested it tapped a general, non mathematical reasoning dimension. Remaining anxiety about this item type is not based on anything in the data analysis but merely reflects a concern with producing questions about ambiguous problems that have unambiguous, defensible answers. An explanation that initially appears to be irrelevant to the test developer may be relevant to an explanation developed by a highly creative examinee.

Perhaps the most interesting result was the comparability of the analytical reasoning tests in the multiple-choice and open-ended formats. Even though the two tests were administered four months apart and students were motivated to do well only on the multiple-choice test, the correlation between the factors representing these two test formats was .93. Only the computer version of the task provided a pictorial representation of the problem; had the multiple-choice version also provided this visual aid, the correlation between formats might have been even higher. Thus, the open-ended version does not appear to tap any significantly different new dimension. These data, combined with the test developers' analysis of the processes needed to answer the multiple-choice questions, suggest that the current items do indeed require the student to generate an answer; the multiple-choice questions simply confirm that the answer generated by the student was correct. The student's or the public's perception of the task may be different in the open-ended computer format, but it is possible that the actual processes tapped may be nearly identical across formats. This result cannot be generalized to other item types where format differences could result in assessment of different underlying constructs.

Although there was evidence that logical reasoning items, analytical reasoning items, and analysis of explanations items were reliably assessing

somewhat different constructs, a GRE general test with three different analytical subscores might prove cumbersome. Further work is needed to produce a coherent analytical score that is distinct from the current verbal and quantitative scores.

## References


Bennett, R. E., & Kaplan, R. (1990). Developing and evaluating a machine-scorable, constructed-response item type for divergent thinking. Proposal submitted to GRE Research Committee.

Bentler, P. M. (1985). Theory and Implementation of EQS: A structural equation program. Los Angeles: BMDP Statistical Software.

Bridgeman, B. (1991). A comparison of open-ended and multiple-choice question formats for the quantitative section of the Graduate Record Examination. (GRE Research Report No. 88-13) Princeton, NJ: Educacional Testing Service.

Carlton, S. T. (1987). Logical and critical thinking. Princeton, NJ: Educational Testing Service.

Emmerich, W., Enright, M. K., Rock, D. A., & Tucker, C. (1991). The development, investigation, and evaluation of new item types for the GRE Analytical measure (GRE Board Professional Report No. 87-09P) Princeton, NJ: Educational Testing Service.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problems solving. Applied Psychological Measurement, 17, 11-29.

Jöreskog, K., & Sörbom, D. (1985). LISREL 6: An analysis of linear structural relationships by the method of maximum likelihood. Mooresville, IN: Scientific Software.

Kingston, N. M., & Dorans, N. J. (1982). The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory (GRE Board Professional Report No. 79-12bP, ETS Research Report 82-88). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & McKinley, R. L. (1988, April). Assessing the structure of the GRE General Test using confirmatory multidimensional item response theory. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First and higher order factor models and their invariance across groups. Psychological Bulletin, 97, 562-582.

Martinez, M. (in preparation). Figural response assessment: System development and pilot research in cell and molecular biology.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. Journal of Higher Education, 58, 658-682.

Powers, D. E., & Swinton, S. S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. Applied Psychological Measurement, 5(2), 141-158.

Rock, D. A., Bennett, R. E., & Jirele, T. (1988). Factor structure of the Graduate Record Examinations General Test in handicapped and nonhandicapped groups. Journal of Applied Psychology, 73, 383-392.

Schaeffer, G. A., & Kingston, N. M. (1988). Strength of the analytical factor of the GRE General Test in several subgroups: A full-information factor analysis approach (GRE Board Professional Report No. 86-7P, ETS Research Report 85-5. Princeton, NJ: Educational Testing Service.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Machine-scorable complex constructed response quantitative items: Agreement between expert system and human raters' scores (GRE Board Professional Report No. 88-07). Princeton, NJ: Educational Testing Service.

Swinton, S. S., & Powers D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. Journal of Educational Psychology, 75, 104-115.

Swinton, S. S., Wild, C. L., & Wallmark, M. (1983). Investigation of practice effects on item types in the Graduate Record Examination Aptitude Test. (GRE Board Professional Report No. 80-1cP). Princeton, NJ: Educational Testing Service.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Wah, D. M., & Robinson, D. S. (1990). Examinee and score trends for the GRE General Test: 1977-78, 1982-83, 1986-87, and 1987-88. Princeton, NJ: Educational Testing Service.

Ward, W. C., Carlson, S. B., & Woisetschlaeger, E. (1983). Ill-structured problems as multiple-choice items (GRE Board Professional Report No. 81-18P). Princeton, NJ: Educational Testing Service.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17, 11-29.

Wilson, K. M. (1985). The relationship of GRE General Test item-type part scores to undergraduate grades (GRE Board Professional Report No. 81-22P, ETS Research Report 84-38). Princeton, NJ: Educational Testing Service.

## Appendix A

## Analytical Reasoning Questions

ANALYTICAL REASONING
Instructions and Practice Question

Analytical reasoning questions test the ability to understand a given structure of arbitrary relationships among fictitious persons, places, things, or events, and to deduce new information from the relationships given. Each item in this section consists of (1) a set of about three to seven related statements or conditions (and sometimes other explanatory material), (2) a set of symbols to be arranged (e.g., people's names, cars, city names), (3) a work space where the symbols may be arranged, and (4) a problem statement that asks for a particular arrangement of the symbols.

Press **ENTER** to continue

---

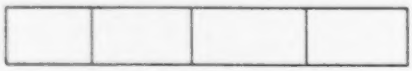
Look at the sample problem presented below:

**Conditions**
You are given four boxes numbered 1, 2, 3, and 4.
The triangle must go in box 2.
Each symbol must go in one and only one box.

**Symbol Storage** **Solution Storage**

○ ☼ ▲ ▬
B S T R

**Work Space**




**Sample Problem** : Place one symbol (ball, star, triangle, or rectangle) in each box.

---

Note that each symbol is identified by a letter. Each box in the work space may be highlighted in turn by pressing the **Space Bar**. Try it now. The Conditions state that the triangle must be in box 2, so when you have box 2 highlighted type the letter **T** and the triangle will move into that box. Then highlight box 1 and type **B**. Highlight box 3 and type **S**. Highlight box 4 and type **R**.

---

Suppose the conditions had required you to place the star in box 1 but you had mistakenly placed it in box 3. Simply use the **Space Bar** to highlight box 1 and type **S**. Do it now. Note that the star replaced the ball and the ball was returned to the Symbol Storage area. Now move the highlight to box 3 and type **B**. When your answer is complete, press **ENTER**. Press **ENTER** again to confirm your answer.

(Other practice items were included in the computer presentation; they are omitted here.)

Now you are ready to start the actual test.

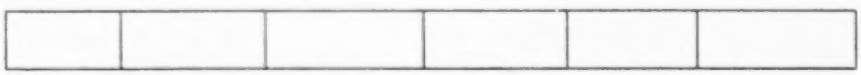
An organist is arranging to judge the playing of original compositions by six student organists--R, S, T, U, V, and W. She will hear one student play each day from Monday through Saturday. She must schedule the auditions for the students according to the following conditions:

R must play earlier in the week than W.
S must play on Thursday.
T must play on the day immediately before or immediately after the day on which U plays.
V cannot play on Tuesday.

**Symbol Storage** R S T U V W

**Solution Storage**

**Work Space**



**Monday** **Tuesday** **Wednesday** **Thursday** **Friday** **Saturday**

---

**Problem 1:** If V plays on the day immediately before T plays, make a possible schedule of auditions. (Answer: VTUSRW)

---

**Problem 2:** If R must play on the day immediately after the day on which V plays, make a possible schedule of auditions. (Answer: VRWSTU or VRWSUT)

---

**Problem 3:** Make another schedule that is different from your responses to Problem 1 and Problem 2. (Answer: 16 correct answers)

---

**Problem 4:** The organist could schedule any of the following to play on a day immediately before or after the day on which T plays EXCEPT

(A) R
(B) S
(C) U
(D) V
(E) W

Answer: E

## Appendix B

## Pattern Identification Questions

## Pattern Identification
## Instructions and Practice Question

In this test, a number series is composed of exactly seven whole numbers (positive integers).

Example: 2 4 6 8 10 12 14

Each number in a series except the first (leftmost) is calculated from the number preceding (to the immediate left) by applying a series rule. For the example above the rule is: "Start with the number in any position, then add 2."
This rule can be represented as follows:

| 2 | 4 | 6 | 8 | 10 | 12 | 14 |
|---|---|---|---|---|---|---|
| | (+2) | (+2) | (+2) | (+2) | (+2) | (+2) |

Enter this selection rule by following the directions at the bottom of the screen.

---

Start with the number in any position, then

| Pick one | | | Add | |
|---|---|---|---|---|
| Add | 1 | | Subtract | 2 |
| Subtract | 2 | then | Multiply by | 3 |
| Multiply by | 3 | | Divide by | 4 |
| Divide by | 4 | | STOP | |

---

Move the selection box by pressing the **Space Bar**. When the box highlights "Add," press **ENTER**. (After response, highlight box moves automatically to top of first column of numbers).

---

Now press the **Space Bar** to move the selection box to the number 2, then press **ENTER**. (highlight box moves to top of second column of operations)

---

No further calculation is required for this problem. Move the selection box to STOP and press **ENTER**.

---

Press **ENTER** again to confirm your choice or press **Space Bar** to erase your response and start over.

A different example of a number series is the following:

```
1       3       7       15       31       63       127
 (x2,+1) (x2,+1) (x2,+1)  (x2,+1)  (x2,+1)  (x2,+1)
```

For this example an applicable series rule is the following: "The number is calculated by multiplying the preceding number by 2 and then adding 1 to the product."

---

Start with the number in any position, then

**Pick one**

| | | | | |
|---|---|---|---|---|
| Add | 1 | | Add | 1 |
| Subtract | 2 | then | Subtract | 2 |
| Multiply by | 3 | | Multiply by | 3 |
| Divide by | 4 | | Divide by | 4 |
| | | | STOP | |

---

Press the **Space Bar** until the selection box is on "Multiply by," then press **ENTER**.

---

Now move the selection box to "2" and press **ENTER**.

---

Now move the selection box to "Add" and press **ENTER**.

---

Now move the selection box to "1" and press **ENTER**. Press **ENTER** again to confirm your choice or press **Space Bar** to erase your response and start over.

In each number thus far a single formula has been used to calculate the numbers. However, an applicable series rule may use more than one formula, in which case the series rule must conform to one of the two patterns described below.

Pattern--One rule for: 2nd, 4th, and 6th
Different rule for: 3rd, 5th, and 7th

Example-- 2 4 3 6 5 10 9
(x2) (-1) (x2) (-1) (x2) (-1)

Pattern--One rule for: 2nd, 3rd, 4th
Different rule for: 5th, 6th, 7th

Example-- 5 8 11 14 11 8 5
(+3) (+3) (+3) (-3) (-3) (-3)

Try this example.

```
5      8      11      14       27       53       105
  (+3)   (+3)    (+3)   (x2,-1)  (x2,-1)  (x2,-1)
```

| Choose Pattern | Examples |
|---|---|
| Same rule for all | 2 4 6 8 10 12 14<br>(+2) (+2) (+2) (+2) (+2) (+2) |
| One rule for: 2nd, 4th, 6th<br>Different rule for: 3rd, 5th 7th | 2 4 3 6 5 10 9<br>(x2) (-1) (x2) (-1) (x2) (-1) |
| One rule for: 2nd, 3rd, 4th<br>Different rule for: 5th, 6th, 7th | 5 8 1 14 11 8 5<br>(+3) (+3) (+3) (-3) (-3) (-3) |

First, move the selection box to the last choice, then press **ENTER**.

```
5      8      11      14       27       53       105
  (+3)   (+3)   (+3)   (x2,-1)  (x2,-1)  (x2,-1)
```

---

Start with the number in position 1, 2, or 3, then
Pick one

| | | | | |
|---|---|---|---|---|
| Add | 1 | | Add | 1 |
| Subtract | 2 | then | Subtract | 2 |
| Multiply by | 3 | | Multiply by | 3 |
| Divide by | 4 | | Divide by | 4 |
| | | | STOP | |

---

Note that you are now asked for the rule stating what to do with the number in position 1, 2, or 3. You want to add 3, so move the selection box to "Add" and press **ENTER**.

---

Now move the selection box to "3" and press **ENTER**.

---

Now move the selection box to "STOP" and press **ENTER**.

---

```
5      8      11      14       27       53       105
  (+3)   (+3)    (+3)   (x2,-1)  (x2,-1)  (x2,-1)
```

---

Start with the number in position 4, 5, or 6, then

**Pick one**

| | | | | |
|---|---|---|---|---|
| Add | 1 | | Add | 1 |
| Subtract | 2 | then | Subtract | 2 |
| Multiply by | 3 | | Multiply by | 3 |
| Divide by | 4 | | Divide by | 4 |
| | | | STOP | |

---

Note that you are now asked for the rule stating what to do with the number in position 4, 5, or 6. Move the selection box to "Multiply by" then press **ENTER**.

---

Select "2" and press **ENTER**. (highlight box moves to second column of operations)

---

Select "Subtract" and press **ENTER**.

---0

Select "1" and press **ENTER**.

Now you are ready to begin the actual test. You will have three minutes to complete each problem. The clock in the corner of the screen tells you how much time you have left.

2 4 3 6 5 10 9

| Choose Pattern | Examples |
|---|---|
| Same rule for all | 2 4 6 8 10 12 14<br>(+2) (+2) (+2) (+2) (+2) (+2) |
| One rule for: 2nd, 4th, 6th<br>Different rule for: 3rd, 5th 7th | 2 4 3 6 5 10 9<br>(x2) (-1) (x2) (-1) (x2) (-1) |
| One rule for: 2nd, 3rd, 4th<br>Different rule for: 5th, 6th, 7th | 5 8 1 14 11 8 5<br>(+3) (+3) (+3) (-3) (-3) (-3) |

First, move the selection box to your choice, then press **ENTER**.

---

PROGRAMMER NOTES

If "Same rule for all" is selected, put this message at the top of the rule statement box:

Start with the number in any position, then

If "One rule for: 2nd, 4th, 6th..." is selected, put this message at the top of the rule statement box:

Start with the number in position 1, 3, or 5, then
(after response to first screen is complete)
Now start with the number in position 2, 4, or 6, then

If "One rule for 2nd, 3rd, 4th..." is selected, put this message at the top of the rule statement box:

Start with the number in position 1, 2, or 3, then
(after response to first screen is complete)
Start with the number in position 4, 5, or 6, then

## PATTERN IDENTIFICATION TEST QUESTIONS

1. 2 4 3 6 5 10 9

2. 3 4 6 10 18 34 66

3. 8 3 6 2 4 1 2

4. 2 3 6 15 27 51 99

5. 5 6 8 12 14 24 26

6. 4 5 7 11 19 35 67

7. 2 3 5 9 17 33 65

8. 6 8 24 32 96 128 384

## Appendix C

## Analysis of Explanations Questions

Set 1

Situation: In an attempt to end the theft of books from Parkman University Library, Elenora Johnson, the chief librarian, initiated a stringent inspection program at the beginning of the fall term. At the library entrance, Johnson posted inspectors to check that each library book leaving the building had a checkout slip bearing the call number of the book, its due date, and the borrower's identification number. The library retained a carbon copy of this slip as its only record that the book had been checked out. Johnson ordered the inspectors to search for concealed library books in attaché cases, bookbags, and all other containers large enough to hold a book. Since no new personnel could be hired, all library personnel took turns serving as inspectors, though many complained of their embarrassment in conducting the searches.

Result: During that term Margaret Zimmer stole twenty-five library books.

Answer Key: A-inconsistent; B-deducible; C-relevant; D-not relevant.

1. Zimmer stole the books before the inspection system began. (Answer: A)

2. The windows in the library could not be opened. (Answer: C)

3. During that term, if Zimmer carried a bookbag out of the library entrance door during regular hours, an inspector was supposed to check it. (Answer: B)

4. The doors to the library fire escapes are equipped with alarm bells set off by opening the doors. (Answer: C)

5. The library had at one time kept two carbon copies of each checkout slip. (Answer: D)

Copyright © 1977 by Educational Testing Service. All rights reserved.
Reproduced by permission.

For "A" answer key

Is the statement inconsistent with, or contradictory to, something in the fact situation, the result, or both together?

**YES** **NO**

If yes, score correct and stop. If no, score incorrect and stop.

---

For "B" answer key

Is the statement inconsistent with, or contradictory to, something in the fact situation, the result, or both together?

**YES** **NO**

If yes, score incorrect and stop. If no, score correct and continue:

Does the statement have to be true if the fact situation and result are as stated? That is, is the statement deducible from something in the fact situation?

**YES** **NO**

If yes, score correct and stop. If no, score incorrect and stop.

---

For "C" answer key

Is the statement inconsistent with, or contradictory to, something in the fact situation, the result, or both together?

**YES** **NO**

If yes, score incorrect and stop. If no, score correct and continue:

Does the statement have to be true if the fact situation and result are as stated? That is, is the statement deducible from something in the fact situation?

**YES** **NO**

If yes, score incorrect and stop. If no, score correct and continue:

Does the statement either support or weaken a possible explanation of the result? That is, is the statement relevant to an explanation of the result?

**YES** **NO**

If yes, score correct and stop. If no, score incorrect and stop.

For "D" answer key

Is the statement inconsistent with, or contradictory to, something in the fact situation, the result, or both together?

**YES** **NO**

If yes, score incorrect and stop. If no, score correct and continue:

Does the statement have to be true if the fact situation and result are as stated? That is, is the statement deducible from something in the fact situation?

**YES** **NO**

If yes, score incorrect and stop. If no, score correct and continue:

Does the statement either support or weaken a possible explanation of the result? That is, is the statement relevant to an explanation of the result?

**YES** **NO**

If yes, score incorrect and continue: If no, score correct and stop.

Briefly explain the result, showing how the statement supports or weakens your explanation. You may use outline form or incomplete sentences, but your answer may not exceed the five lines provided.

_____

_____

_____

_____

_____

Press the F1 key if you cannot think of an explanation.

Press the F2 when you are finished.

Appendix D

Posttest Questions

*Percentage selecting each option is in parentheses.*

GRE COMPUTER TEST

POSTTEST QUESTIONNAIRE

GRE Registration Number_____

The test you just took had five sections. The following questions refer to these sections by number:

1. Quantitative comparisons (the quantity in A is greater...)
2. Regular math ( $(2x)^3+(3y)^2=$ )
3. Analytical reasoning (If V plays on the day immediately before T plays, make a possible schedule)
4. Number series (2 4 6 8...)
5. Analysis of explanations (Is the statement inconsistent with something in the fact situation..Yes..No)

1. For each section of the test, indicate whether you thought that section was too easy, about right, or too difficult.

| | (Circle one number for each section) | | |
|---|---|---|---|
| Section | Too easy | About right | Too difficult |
| 1 | 1 (11) | 2 (76) | 3 (13) |
| 2 | 1 (10) | 2 (74) | 3 (16) |
| 3 | 1 (22) | 2 (65) | 3 (13) |
| 4 | 1 (5) | 2 (44) | 3 (52) |
| 5 | 1 (21) | 2 (61) | 3 (19) |

2. For each section of the test, indicate whether time limits were too short (couldn't finish), about right (just enough time to finish), or too long (lots of time left over).

| | (Circle one number for each section) | | |
|---|---|---|---|
| Section | Too short | About right | Too long |
| 1 | 1 (18) | 2 (68) | 3 (14) |
| 2 | 1 (38) | 2 (58) | 3 (4) |
| 3 | 1 (27) | 2 (65) | 3 (9) |
| 4 | 1 (53) | 2 (43) | 3 (4) |
| 5 | 1 (5) | 2 (60) | 3 (35) |

3. For each section, indicate whether the directions were easy to understand or hard to understand.

| Section | (Circle one number for each section) Easy directions | Hard directions |
|---|---|---|
| 1 | 1 (91) | 2 (9) |
| 2 | 1 (91) | 2 (9) |
| 3 | 1 (86) | 2 (14) |
| 4 | 1 (51) | 2 (49) |
| 5 | 1 (68) | 2 (32) |

Describe any problems that you had with the directions.

_____

_____

_____

4. In some sections of the test there were separate time limits for individual questions; in other sections of the test you had a fixed amount of time for the whole section. Assuming that the number of minutes per question were the same with either timing strategy, which do you prefer?

(Circle your choice)

| Time each question | Time entire section |
|---|---|
| 1 (33) | 2 (67) |

5. In the section of the test where you could make the clock disappear with the F1 key, did you ever use that key? (Circle your choice)

| Yes | No |
|---|---|
| 1 (8) | 2 (92) |

6. Did you wish that you could make the clock disappear on other sections of the test? (Circle your choice)

| Yes | No |
|---|---|
| 1 (24) | 2 (76) |

7. Think back to when you worked on quantitative items on the regular GRE in October. Did you ever try to work backwards from the answer choices or use any other strategy that you could not use on the open-ended quantitative questions on the computer test that you just completed?

(Circle your choice)

| Yes | No |
|---|---|
| 1 (88) | 2 (12) |

8. Which kind of quantitative test would you rather take, the multiple-choice type on the regular GRE or the open-ended type on the computer test?

(Circle your choice)

| Multiple-choice | Open-ended | No difference |
|---|---|---|
| 1 (81) | 2 (11) | 3 (8) |

9. Which kind of quantitative test do you think is a fairer indicator of your quantitative ability, the multiple-choice type on the regular GRE or the open-ended type on the computer test? (Circle your choice)

| Multiple-choice | Open-ended | No difference |
|---|---|---|
| 1 (43) | 2 (41) | 3 (16) |

10. Think back to when you worked on analytical reasoning items on the regular GRE in October. Did you ever try to work backwards from the answer choices or use any other strategy that you could not use in the analytical reasoning items on the computer test? (Circle your choice)

| Yes | No |
|---|---|
| 1 (69) | 2 (31) |

11. Which kind of analytical reasoning test would you rather take, the multiple-choice type on the regular GRE or the open-ended type on the computer test? (Circle your choice)

| Multiple-choice | Open-ended | No difference |
|---|---|---|
| 1 (52) | 2 (32) | 3 (16) |

12. Which kind of analytical reasoning test do you think is a fairer indicator of your reasoning ability, the multiple-choice type on the regular GRE or the open-ended type on the computer test? (Circle your choice)

| Multiple-choice | Open-ended | No difference |
|---|---|---|
| 1 (33) | 2 (35) | 3 (33) |

13. In the past two years, how often have you used a computer for word processing? (Circle your choice)

| Never | 1-5 times | More than 5 times |
|---|---|---|
| 1 (18) | 2 (28) | 3 (79) |

14. In the past two years, how often have you used a computer for anything other than word processing? (Circle your choice)

| Never | 1-5 times | More than 5 times |
|---|---|---|
| 1 (18) | 2 (28) | 3 (55) |

15. When you have to write a paper, how do you usually do it? (Circle your choice)

| Typewriter | Computer | Pencil or pen |
|---|---|---|
| 1 (17) | 2 (64) | 3 (19) |

If you have any additional comments or suggestions for improving the test, please use the space below or the opposite side of this page.

54020 08911 • Y13M.7 • 209069